Historic, archived document

Do not assume content reflects current scientific knowledge, policies, or practices.

UNITED STATES
DEPARTMENT OF AGRICULTURE
DEPARTMENT CIRCULAR 307
Washington, D. C. March 31, 1924

A METHOD OF TESTING FARM-MANAGEMENT AND COST-OF-PRODUCTION DATA FOR VALIDITY OF CONCLUSIONS.

H. R. TOLLEY,
In Charge, Division of Farm Management,
and
S. W. MENDUM,
Junior Economist.
Bureau of Agricultural Economics

CONTENTS.

	Page.
Two-way frequency table and scatter diagram	2
Correlation coefficient	6
Regression line and regression coefficient	8
Application of the coefficients	12

The purpose of farm-business analysis is to measure the relative importance of the several factors entering into production, and to account for the obvious differences in results arrived at by different men under different conditions of time, and place and with different types of farms, etc.

Starting with general ideas of the normal relationships existing in familiar situations, the investigator assembles his data and interprets them in the light of his experience, through the process of grouping and averaging. Grouping and averaging give a definite impression, which usually supports the conclusions on the basis of which the groupings of data are made or suggests modifications in them. Where relations between factors are intimate and simple, grouping and averaging give satisfactory results, quantitatively as well as qualitatively.

Relatively few of the relationships of economic data, however, are sufficiently close and free from complication by other significant influences to yield to the initial process of grouping and averaging alone. Recognizing this, it is customary to make numerous qualifications to any general conclusions apparently warranted by summary tables.

76109°—24——1

This circular aims to show the application to farm-management and cost-of-prodution data,[1] of methods of analysis developed in other lines of research, for the specific purpose of giving quantitative expression to the relationships of the widely variable characteristics so frequently observed.

The devices employed are, essentially, checks on the validity of conclusions which seem to be warranted by the body of data under study, as a result of the customary steps of analysis of which grouping and averaging are the first. They may also be considered as measures or checks on the amount of divergence of the data at hand from generally accepted principles. In either case, the process of analysis throws a spot light on the entire situation and factors often otherwise obscure are brought out.

TWO-WAY FREQUENCY TABLE AND SCATTER DIAGRAM.[2]

The first and simplest of these analytical devices and that giving the maximum return for the minimum of effort, is the double-frequency table and its graphic representation, the "dot chart," or

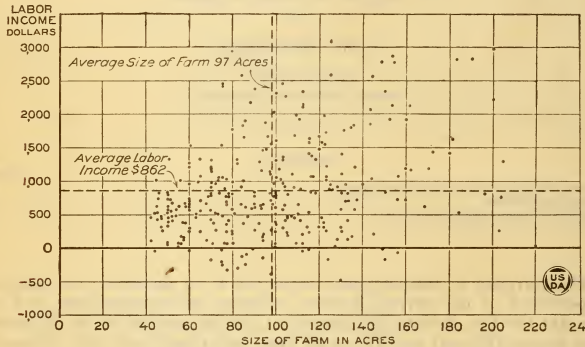

FIG. 1.—*Relation of size of farm to labor income.*—Each dot represents the labor income of a farm of the size indicated by its position on the horizontal scale. The figures were obtained for 320 farms in Chester County, Pa., for the year 1911. Note the wide differences in labor incomes of farms of practically the same size.

"scatter diagram." This device is not unknown to farm-management workers, though apparently it has not been regularly applied as a check, but has been used chiefly as a method of graphic presentation of results instead of a step in analysis.

Take, for example, the proposition that labor income increases as size of farm increases, conceded to be generally true. In this case, we have a simple variable (number of acres in farm) measured in terms of another variable (labor income), which is the resultant of a large number of factors, the simple variable being only one of them. The question is this: Is size of farm as significant as the general

[1] The illustrations herewith presented were worked out by members of the Bureau of Agriculture Economics as exercises in a course in The Theory of Statistics, given by Mr. Tolley during the school year 1922-23 in the Graduate School, Department of Agriculture.

[2] Also called correlation table and correlation chart, respectively.

proposition supported by the usual frequency table might lead us to believe? The question comes up, because some large farms make no money and some small farms make a lot of money. Compare Tables 1 and 2, which are summaries of the same body of data. Both tables show the average labor incomes of farmers on farms of different sizes in Chester County, Pa., in 1911. Table 1 is the ordinary frequency table, and indicates a direct and close relationship

TABLE 1.—*Average labor incomes on farms of different sizes (320 farms)—Chester County, Pa., 1911.*

Size of farms.	Number of farmers.	Average labor income.
50 acres and less	27	$547
51 to 80 acres	94	666
81 to 110 acres	96	859
111 to 140 acres	65	950
141 to 170 acres	24	1,394
171 acres and over	14	1,488
All	320	862

between size of farm and labor income. Table 2 shows the number of farmers with labor incomes of different sizes within each group, and indicates that the relationship is not close or necessary, as the labor incomes of some of the men on the largest farms are less than the labor incomes of some of the men on the smallest farms.

TABLE 2.—*Relation between size of farm and labor income—Chester County, Pa., 1911.*

	Number of farmers.	Number with specified labor incomes.						Average.
		Operated at a loss.	$0 to $500.	$501 to $1,000.	$1,001 to $1,500.	$1,501 to $2,000.	$2,001 and over.	
50 acres and less	27	1	12	12	2	.....	.....	$547
51 to 80 acres	94	7	29	39	13	3	3	666
81 to 110 acres	96	8	22	33	15	11	7	859
111 to 140 acres	65	4	17	17	12	10	5	950
141 to 170 acres	24	3	2	3	5	3	8	1,394
171 acres and over	14	.....	3	2	3	1	5	1,488
All	320	23	85	106	50	28	28	862

Notice that Table 2 shows all the figures shown in Table 1, but gives as additional information the differences observed within the groups. Figure 1 graphically shows the detail for the 320 farms.

Table 3 indicates that there is a close relationship between farm receipts and farm expenses on 238 farms operated by white owners in Sumter County, Ga., in 1913. The distribution within the groups is narrow, and the indication of the ordinary one-way table is supported by the detail. Figure 2 is the graphic presentation of the figures in Table 3.

In making the scatter diagram or correlation chart the position of any dot representing an observation is determined by reference to its numerical value on two scales—one horizontal, the other vertical—at the intersection of the two values projected. It is customary to measure the independent, or causal variable (size of farm in fig. 1), on

TABLE 3.—*Relation between farm expenses and farm receipts (238 farms operated by white owners)—Sumter County, Ga., 1913.*

Farm expenses (class limits).	Number of farms.	Number of farms with specified farm receipts.				
		$0 to $949.	$950 to $1,949.	$1,950 to $2,949.	$2,950 to $3,949.	$3,950 to $4,949.
$0 to $449	22	21	1			
$450 to $949	41	16	24	1		
$950 to $1,449	39		26	13		
$1,450 to $1,949	36		2	20	14	
$1,950 to $2,449	30			7	20	3
$2,450 to $2,949	15			2	5	7
$2,950 to $3,449	15				1	6
$3,450 to $3,949	7					2
$3,950 to $4,449	5					
$4,450 to $4,949	7					
$4,950 to $5,449	10					
$5,450 to $5,949	6					
$5,950 to $6,449	2					
$6,450 to $6,949	2					
$6,950 to $7,949	1					
All	238	37	53	43	40	18

Farm expenses (class limits).	Number of farms.	Number of farms with specified farm receipts.					Average farm receipts.
		$4,950 to $5,949.	$5,950 to $6,949.	$6,950 to $7,949.	$7,950 to $8,949.	$8,950 to $9,949.	
$0 to $449	22						$520
$450 to $949	41						1,100
$950 to $1,449	39						1,800
$1,450 to $1,949	36						2,740
$1,950 to $2,449	30						3,270
$2,450 to $2,949	15	1					3,890
$2,950 to $3,449	15	6		1	1		5,370
$3,450 to $3,949	7	4			1		5,640
$3,950 to $4,449	5	4	1				5,720
$4,450 to $4,949	7		1	5	1		7,600
$4,950 to $5,449	10	2	2	2	2	2	7,440
$5,450 to $5,949	6	1	1	1	2	1	7,680
$5,950 to $6,449	2				1	1	8,750
$6,450 to $6,949	2					2	9,200
$6,950 to $7,949	1					1	9,800
All	238	18	5	9	8	7	3,150

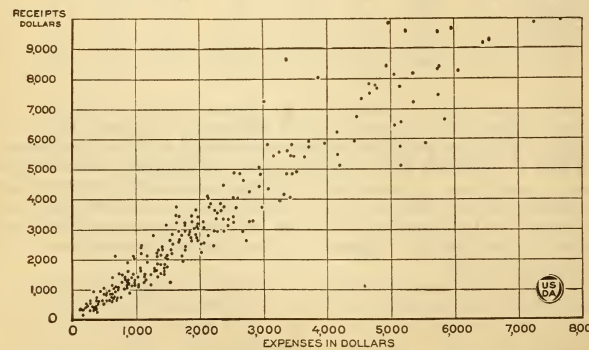

FIG. 2.—*Relation of farm expenses to farm receipts.*—Farm expenses and farm receipts are closely correlated. These figures are from the records of 238 farms operated by white owners in Sumter County, Ga., in 1913.

the horizontal scale, and the dependent variable, or resultant (labor income), on the vertical scale. The scales should provide for the full range of the observations in any particular group. Whether or not both scales are the same length, making a square chart, is a matter of convenience, but usually the square chart will be found most satisfactory.

TABLE 4.—*Relation between the cost per pound of gain and the loss or profit per head on 865 droves of steers fattened in Illinois, Iowa, and Nebraska, winters 1918–19 and 1919–20.*

Cost per pound gain (cents).	Number of droves.	Number of droves with specified loss or profit per head.													
		−$135.99 to −$128.1	−$127.99 to −$120.	−$119.99 to −$112.	−$111.99 to −$104.	−$103.99 to −$96.	−$95.99 to −$88.	−$87.99 to −$80.	−$79.99 to −$72.	−$71.99 to −$64.	−$63.99 to −$56.	−$55.99 to −$48.	−$47.99 to −$40.	−$39.99 to −$32.	−$31.99 to −$24.
6 to 9.9	2														
10 to 13.9	39														
14 to 17.9	92														1
18 to 21.9	140										1		1	3	4
22 to 25.9	154											1	2	9	12
26 to 29.9	133										2	4	8	15	22
30 to 33.9	110									1	1	6	17	14	23
34 to 37.9	81							2	2	2	8	4	10	18	9
38 to 41.9	38								1		2	6	8	7	2
42 to 45.9	28							2	1	4	4	2	5	1	2
46 to 49.9	22					1	1		1	3	3	2	1	6	2
50 to 53.9	16			1		1	1	1		1	3	1	1	4	1
54 to 57.9	3										1				
58 to 61.9	3	1										1	1		
62 to 65.9															
66 to 69.9	1											1			
70 to 73.9	1											1			
74 to 77.9	1										1				
78 to 81.9															
82 to 85.9															
86 to 89.9	1													1	
Total	865	1		1		2	2	5	5	11	26	29	54	78	78

Cost per pound gain (cents).	Number of droves.	Number of droves with specified loss or profit per head.												Average profit or loss per head.[1]
		−$23.99 to −$16.	−$15.99 to −$8.	−$7.49 to $0.	$0 to $7.99.	$8 to $15.99.	$16 to $23.99.	$24 to $31.99.	$32 to $39.99.	$40 to $47.99.	$48 to $55.99.	$56 to $63.99.	$64 to $71.99.	$72 to $79.99.	
6 to 9.9	2							1			1				$42.00
10 to 13.9	39	1	5	4	5	6	5	4	4	3	1	1			15.13
14 to 17.9	92	1	6	15	16	24	8	7	4	6	1	2		1	13.04
18 to 21.9	140	7	23	32	24	13	17	8	4	2	1				1.29
22 to 25.9	154	27	32	27	16	13	9	4	2						−8.21
26 to 29.9	133	19	13	17	11	10	5	6	1						−14.63
30 to 33.9	110	15	7	16	3	4	2		1						−24.25
34 to 37.9	81	11	3	4	2	3	2	1							−32.54
38 to 41.9	38	4	6	1				1							−34.21
42 to 45.9	28	3		3	1										−44.43
46 to 49.9	22	1		1											−49.27
50 to 53.9	16			1											−56.25
54 to 57.9	3	2													−34.00
58 to 61.9	3														−76.67
62 to 65.9															
66 to 69.9	1														−54.00
70 to 73.9	1														−34.00
74 to 77.9	1														−58.00
78 to 81.9															
82 to 85.9															
86 to 89.9	1														−38.00
Total	865	91	95	121	78	73	48	32	16	11	4	3		1	−13.32

$r = -0.658$.

$b = -1.78$, i. e., on the average, an increase of 1 cent in the cost per pound of gain was accompanied by a decrease of $1.78 in profit per head.

[1] A minus sign indicates loss.

In Figure 1 the average values of each variable are drawn, dividing the chart into four quadrants. This helps to visualize the relation of the observations to the averages. The lack of concentration of the dots in any particular part of the chart shows that there is not a close relationship between the variables. In Figure 2 the class limits are drawn, instead of the average values.

In Tables 2 and 3 and Figures 1 and 2 the variables both increase. This tendency for both to increase together is called direct, or positive, correlation. Table 4 and Figure 4 illustrate inverse or negative correlation; that is, there is a tendency for profits per head to decrease as cost per pound of gain increases. It is evident that the relationship between these two variables is closer than between those of Table 2, but not so close as between those of Table 3.

CORRELATION COEFFICIENT.

The process of grouping and averaging, whether by the one-way or the two-way (double) frequency table and scatter diagram, gives only a *qualitative* answer to the question of relationships between the variables. A *quantitative* measure of the degree of relationship is needed. The correlation coefficient is such a measure, and is practically the only measure thus far devised for determining to what extent changes from the average in one quantity are simultaneous with and proportionate to changes from the average in another, and which of two or more causal factors is most closely associated with the resultant.

The method of computing the coefficient of correlation is illustrated in Table 5. The general observation, on the 48 dairy farms studied, was that the higher the annual production of milk per cow the smaller the number of hours of labor required to produce 100 pounds of milk. This question arose: Is the general observation reliable? The coefficient of correlation (r)[3] proves to be -0.58, which, under the circumstances, is fairly satisfactory evidence that the relation is close and negative, and indicates that closer observation of details and a definite purpose on the part of farmers to cut labor to the amount necessary under the best conditions would result in a much higher coefficient. The procedure follows.

Notice that all the figures for the same farm are listed on the same line; there must be no confusion of details.

1. Find the average value of each of the variables (X and Y). In this case the average annual production per cow was 7,220 pounds of milk and the average labor per 100 pounds was 2.67 hours.

2. Find the deviation or difference (x and y) of each of the individual variables from its average. In the case of Dairy No. 9, the deviation of production from the average (x) is +1,460 pounds ($x_9 = X_9 - A_x$), and the deviation of the labor from the average (y) is -1.34 hours ($y_9 = Y_9 - A_y$). The algebraic sign is important; differences below the average are minus ($-$), the rest plus ($+$).

3. Find for each variable the square root of the average of the squares of the deviations; that is, square each x and each y, add each series, divide by the total number of observations, and extract the square root of the quotient. This is the standard deviation (σ). Here the standard deviation of the annual production per cow is.

[3] r is the conventional symbol for "correlation coefficient."

2,081 pounds, and the standard deviation of labor per 100 pounds of milk is 0.976 hour.

4. For each record, find the product of x and y (the deviations of the variables from their respective averages), and find the average of the products. In this case, the sum of the products is -56.44, and the average product (p) -1.18.

5. Divide the average of the products by the products of the two standard deviations to obtain the coefficient of correlation (r), which, in the example, is -0.58.[4]

TABLE 5.—*Calculation of coefficient of correlation between annual production per cow and hours of labor per 100 pounds of milk—48 Wisconsin farms, 1920.*

Farm No.	Annual production, per cow, in thousands of pounds.	Deviation from average.	Squares.	Hours of labor per 100 pounds of milk.	Deviation from average.	Squares.	Products.
	X	x	x^2	Y	y	y^2	xy
9	8.68	1.46	2.13	1.33	−1.34	1.80	−1.96
3	12.15	4.93	24.30	1.49	−1.18	1.39	−5.82
1	13.00	5.78	33.41	2.84	.17	.03	.98
5	9.68	2.46	6.05	1.23	−1.44	2.07	−3.54
12	8.20	.98	.96	1.84	−.83	.69	−.81
6	9.55	2.33	5.43	1.81	−.86	.74	−2.00
8	8.92	1.70	2.89	2.44	−.23	.05	−.39
2	12.60	5.38	28.94	1.64	−1.03	1.06	−5.54
10	8.60	1.38	1.90	1.50	−1.17	1.37	−1.61
13	8.10	.88	.77	1.80	−.87	.76	−.77
4	10.50	3.28	10.76	2.30	−.37	.14	−1.21
15	8.05	.83	.69	1.87	−.80	.64	−.66
11	8.36	1.14	1.30	2.71	.04	0	.05
35	5.94	−1.28	1.64	1.99	−.68	.46	.87
17	8.00	.78	.61	1.87	−.80	.64	−.62
27	6.60	−.62	.38	2.00	−.67	.45	.42
44	4.95	−2.27	5.15	2.95	.28	.08	−.64
25	6.76	−.46	.21	1.75	−.92	.85	.42
31	6.16	−1.06	1.12	2.71	.04	0	−.04
14	8.10	.88	.77	1.85	−.82	.67	−.72
26	6.75	−.47	.22	2.51	−.16	.03	.08
16	8.04	.82	.67	3.43	.76	.58	.62
23	7.21	−.01	0	1.97	−.70	.49	.01
18	7.83	.61	.37	2.10	−.57	.32	−.35
32	6.15	−1.07	1.14	1.90	−.77	.59	.82
30	6.32	−.90	.81	2.14	−.53	.28	.48
19	7.65	.43	.18	2.17	−.50	.25	−.22
43	5.08	−2.14	4.58	4.73	2.06	4.24	−4.41
47	4.18	−3.04	9.24	3.07	.40	.16	−1.22
28	6.48	−.74	.55	2.62	−.05	0	.04
20	7.50	.28	.08	2.59	−.08	.01	−.02
33	6.08	−1.14	1.30	4.96	2.29	5.24	−2.61
40	5.34	−1.88	3.53	2.73	.06	0	−.11
36	5.84	−1.38	1.90	4.41	1.74	3.03	−2.40
34	6.02	−1.20	1.44	2.57	−.10	.01	.12
42	5.22	−2.00	4.00	4.12	1.45	2.10	−2.90
24	6.98	−.24	.06	2.79	.12	.01	−.03
7	9.35	2.13	4.54	3.20	.53	.28	1.13
22	7.29	.07	0	2.54	−.13	.02	−.01
29	6.33	−.89	.79	2.24	−.43	.18	.38
39	5.60	−1.62	2.62	3.37	.70	.49	−1.13
45	4.77	−2.45	6.00	2.90	.23	.05	−.56
46	4.55	−2.67	7.13	3.47	.80	.64	−2.14
38	5.75	−1.47	2.16	3.56	.89	.79	−1.31
41	5.30	−1.92	3.69	3.54	.87	.76	−1.67
48	2.83	−4.39	19.27	5.85	3.18	10.11	−13.96
21	7.43	.21	.04	2.90	.23	.05	.05
37	5.76	−1.46	2.13	3.72	1.05	1.10	−1.53
48 farms	346.53	.....	207.85	128.02	.....	45.70	−56.44
Average	7.22	.....	4.33	2.67	.....	.95	−1.18

$$\sigma_x = 2.081 \qquad \sigma_y = 0.976 \qquad r = \frac{-1.18}{(2.081)\ (0.976)} = -0.58$$

[4] The textbooks contain descriptions of the various methods of calculating the correlation coefficient from correlation tables and of shorter methods than the one used here of calculating it from ungrouped data.

If there is approximately the same number of positive and negative products, and they are of the same size, it will be evident that there is no correlation; and this will be shown by the fact that the coefficient of correlation will be zero, or nearly zero. If high values of the first variable are associated with high values of the second, and low values of the first with low values of the second, most of the products will be *plus;* and the greater their sum, the closer will be the correlation, and the larger will be the coefficient. If a value of one variable below the average is generally associated with a value of the other above the average, the correlation evidently will be negative, and this will be shown by the fact that the sum of the products will be negative, the degree of the correlation and the size of the coefficient depending upon the size of this sum.

The numerical value of the coefficient of correlation (r) is always between $+1$ and -1, $+1$ indicating perfect positive or direct correlation, and -1 indicating perfect negative or inverse correlation, and zero indicating no correlation.

The coefficient of correlation (r) between size of farm and labor income (Table 2) is $+0.32$; between farm expenses and farm receipts (Table 3) $+0.94$; and between cost per pound of gain and profit per head (Table 4), -0.66.

Where several factors, each varying more or less independently, are associated with an observed resultant or dependent variable, the relative importance of the several factors contributing to the resultant is measured by the size of the correlation coefficient, each calculated as just described.

Thus, besides size of farm, as indicated in Table 2, many other factors, such as expenses, receipts, rate of production of cows per farm or per cow, amount of labor, amount of capital, etc. enter into labor income. The correlation of each of these with labor income indicates their relative importance in the body of data being analyzed. (It is quite likely that the correlation of any of the factors named will be different in the data from different areas.) By an extension of the same methods, the combined effect of two or more of these factors on labor income (or on any other factor) may be computed, and the most logical points for attention in reorganization or cost studies be brought out. The arithmetic is extensive but not particularly difficult.

REGRESSION LINE AND REGRESSION COEFFICIENT.

The amount of change in one variable associated with a given change in another is always of interest. Some light on this is obtained by comparing the differences between the group averages, but usually there are irregularities in the amount of change in the dependent variable associated with uniform changes in the independent variable. An expression which will apply to all the groups with the least possible error is needed. The *regression*[5] *line* (graphic) and the *regression coefficient* (numerical) are such expressions.

A *regression line,* or "line of best fit," may be drawn on a correlation chart, or scatter diagram, without more computation than that necessary to determine the position of the intersection of the averages of the two variables. Usually a straight line will serve as well as

[5] The term "regression" was first applied to the inheritance of stature, it being observed that the stature of the sons of tall men tended to regress toward the general mean.

any other. It is drawn through the intersection of the average lines in such a way that half the observations fall on each side of it. The regression lines shown on Figures 3, 4, and 5 are so drawn.

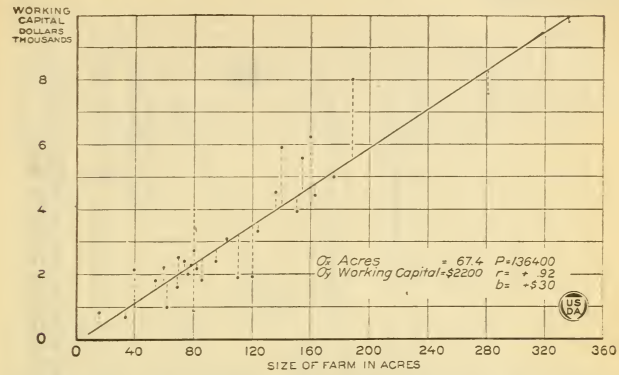

FIG. 3.—*Relation of size of farm to working capital.*—Correlation chart with the regression line drawn according to definition; so that the sum of the squares of the deviations of the observations from the regression line (the squares of the lengths of the dotted lines to the dots) is smaller than they would be from any other straight line. The figures are for 35 farms in Clinton County, Ind., in 1918.

The *regression coefficient* may be computed roughly by dividing the vertical movement of the regression line by the horizontal movement in terms of their respective scales, giving the "slope" of the line, or the amount of change in the dependent variable associated with a unit of change in the independent variable.

The regression coefficient is always related to the correlation coefficient and the standard deviation of the two variables by the formula

$$b = r\frac{\sigma_y}{\sigma_x}$$

where b is the regression coefficient desired, r the coefficient of correlation, σ_y the standard deviation of the dependent variable, and σ_x the standard deviation of the independent variable.

For every set of data, there are two coefficients of regression: The one just defined as $r\frac{\sigma_y}{\sigma_x}$, which is termed the regression of y on x; and one which may be defined as $r\frac{\sigma_x}{\sigma_y}$, which is termed the regression of x on y. Usually, only the first one—the regression of the dependent variable on the independent—is of interest. It is evident from the description of the calculation of the coefficient of correlation that also

$$b = \frac{p}{\sigma^2_x}$$

where p = the average of the products of the deviations, as in Table 5.

For the data shown in Figure 3, $r = +0.92$; $\sigma_y = \$2{,}200$; $\sigma_x = 67.4$ acres and $b = +\$30$. That is, an increase of 1 acre in the size of the farm was accompanied, on the average, by an increase of \$30 in working capital.

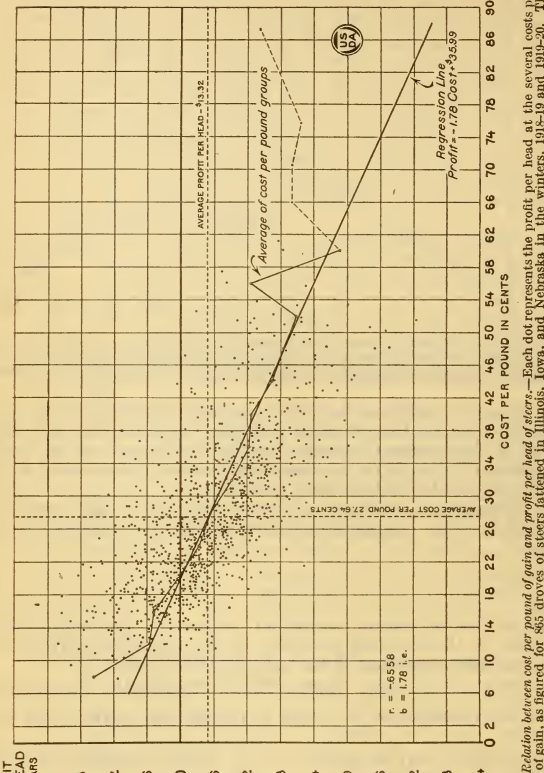

FIG. 4.—*Relation between cost per pound of gain and profit per head of steers.*—Each dot represents the profit per head at the several costs per pound of gain, as figured for 865 droves of steers fattened in Illinois, Iowa, and Nebraska in the winters, 1918–19 and 1919–20. The correlation is reasonably good. Note that the line of group averages is close to the regression line except at the extremes.

Mathematically, the "line of best fit" is a line so drawn that the sum of the squares of the distance of the individual observations (dots) from it is a minimum. The diagonal line in Figure 3 is so drawn. It shows that "on the average," for the 35 owner-operated farms in Clinton County, Ind., in 1918, an increase of 40 acres in size of farm was accompanied by an increase of about \$1,200 in working capital.

The use of a straight line assumes a uniform rate of increase or decrease in the variables throughout the range, but this assumption is not always valid. In Figure 4, for example, the straight regression line fits the data fairly well through the body, that is, up to 52 cents per pound cost of gain, but not beyond. The small number of observations beyond 52 cents suggests that the straight line is really reliable and that some influences not apparent affected the returns, so as to reduce the loss per head. A somewhat different situation is represented by Figure 5, in which the line joining the group averages does not follow the straight regression line closely. Such a regression is called "non-linear regression."

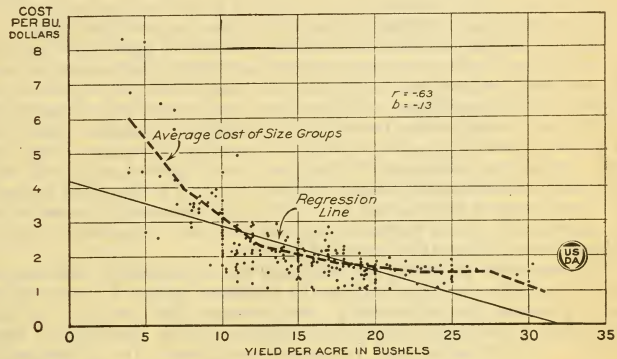

FIG. 5.—*Relation between yield per acre and cost per bushel in wheat production.*—The dots represent the cost per bushel at different yields per acre on 216 farms producing wheat in Oklahoma, Kansas, Nebraska, and Missouri, in 1920. Note that while the correlation coefficient is almost the same as that of the data of Figure 4 the line joining the averages of the yield groups is not so close to the regression line. This is an illustration of "non-linear" regression, and indicates that a different interpretation may be necessary than when the data fall close to the straight line.

TABLE 6.—*Relation between yield of wheat per acre and production cost per bushel on 216 farms (Oklahoma, Kansas, Nebraska, and Missouri) in 1920.*

Yields per acre.	Total number of farms.	Number of farms with specified cost per bushel.						Average cost per bushel.
		$5 and over.	$4 to $4.99.	$3 to $3.99.	$2 to $2.99.	$1 to $1.99.	Under $1.	
Less than 5 bushels..	4	2	2					$5.99
5 to 9.99 bushels.....	23	4	2	12	4	1		3.98
10 to 14.99 bushels....	72		2	6	42	22		2.32
15 to 19.99 bushels....	75				26	49		1.81
20 to 24.99 bushels....	30					30		1.54
25 to 29.99 bushels....	9					9		1.54
30 bushels and over..	3					2	1	1.25
Total.....	216	6	6	18	72	113	1	2.23

For data where the relationship between the two variables is non-linear, the coefficient of regression does not have a direct interpretation as in the example cited above. Likewise, in such cases, the correlation coefficient does not give a true indication of the degree of relationship between the variables. Thus, there are many sets of

farm-management and cost-of-production data which can not be analyzed satisfactorily by the exclusive use of correlation and regression coefficients.

For instance, it is evident, from Table 6 and Figure 5, that the relationship between yield per acre and the cost per bushel of producing winter wheat is non-linear—that is, a straight line does not properly represent the effect of change in yield on the cost of production. The regression line does not follow closely the line joining the average costs of the different yield groups.

The coefficient of regression is -0.13, and while the statement that "on the average" a decrease of 1 bushel in the yield per acre was accompanied by an increase of 21 cents per bushel in the cost of production is still warranted, it does not hold true except on the average, and is not applicable throughout the range of the data.

The size of the correlation coefficient depends upon how closely the individual dots are grouped about the straight regression line on the correlation chart; and, obviously in cases where the regression is nonlinear, the correlation coefficient does not give a true measure of the degree of relationship between the variables. The dots might all lie almost exactly on some curved line, such as is indicated by the lines joining the averages of the groups in Figure 5, and still the coefficient of correlation might be very low. For the data shown, the coefficient of correlation is -0.63, approximately the same as for the data shown in Figure 4, yet the dots are "scattered" less on Figure 5 chart than they are on Figure 4.

The complete problem, in studying the relationship between a pair of variables, is to determine the kind and shape of curve which best represents the relationship, and measure the "scatter" of the individual observations from this curve. Before deciding to rely entirely on correlation and regression coefficients, the investigator should satisfy himself that a straight line represents the relationship between the variables under consideration better than, or at least as well as, any curve.

APPLICATION OF THE COEFFICIENTS.

Having analyzed a set of data, the next thing is to apply the findings to other farms for which only a few of the main factors are readily available. How far one may go in using the rates and coefficients computed from a body of data in estimating results for other farms depends on how well the observed data represent the facts for the other farms, and on the degree of correlation of the factors as shown by the size of the correlation coefficient (r).

If a survey has been carefully made, and if the sample may be assumed to be typical of the area, then the estimates made by using the computations from the observed data ought to apply to other farms not included in the survey with about the same degree of accuracy as was true of the observed data. This is the general basis of research philosophy.

The regression (or estimating) equation is a means of getting a quantitative answer, approximately as accurate as the survey rates and averages, to some of the questions farmers ask. The general form of the regression equation is, according to definition, $y=bx$, but in this form it is not immediately useful because it is in terms of deviations from the averages (A), while definite numbers are wanted.

To fit the general form to the purpose, the values of y $(=Y-A_y)$ and of x $(=X-A_x)$ must be substituted. X and Y are the individual values of the factors observed, or wanted. The equation then becomes: $Y-A_y=b\ (X-A_x)$, or $Y=A_y-bA_x+bX$, the term A_y-bA_x reducing to a single figure (constant) which does not change for the series.

Use the data of Table 5 for illustration. The survey shows that the average production (A_x) of cows in representative herds was 7.22 thousand pounds, and the average hours of labor spent on care of cows (A_y) was 2.67 hours, and the regression coefficient (b) of labor (y) on milk production (x) is -0.272. The constant term A_y-bA_x reduces to 4.63 [equal to $2.67-(-0.272)$ (7.22)], and an *estimated* Y for any farm in the series of Table 5, computed from the equation $Y=4.63-0.272X$, will differ from the *observed* Y, as given in Table 5, by an amount dependent on the amount of deviation of the X and of the Y from the respective averages, but is the closest approximation that can be made. The "fit" is not very close, because the correlation coefficient is not very high, —— only -0.58. In the case of Farm 9, the estimated Y is 2.27 hours per 100 pounds of milk produced, while the observed figure was 1.33 hours. There are five farms the herds of which averaged about 8,000 pounds per cow, from which a labor rate of 2.45 hours would be estimated from the equation; whereas farms 13, 15, 17, 18, and 16 were observed to have used 1.80, 1.87, 1.87, 2.10, and 3.43 hours, respectively. Lacking more definite information, however, the estimated figure 2.45 has a limited use as the best figure obtainable where both factors show such wide variations. A higher correlation coefficient would permit a closer estimate. Feed supply is a more important factor than labor in milk production, so the estimate is perhaps as good as could be expected.

ADDITIONAL COPIES
OF THIS PUBLICATION MAY BE PROCURED FROM
THE SUPERINTENDENT OF DOCUMENTS
GOVERNMENT PRINTING OFFICE
WASHINGTON, D. C.
AT
5 CENTS PER COPY

▽